Deployment of Privacy-Preserving Machine Learning for Political Polling in the 2024 Presidential Election

Sam Buxbaum

Lucas M. Tassis, Lucas Boschelli, Giovanni Comarela, Mayank Varia, Mark Crovella, Dino P. Christenson

BOSTON UNIVERSITY

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

WashU

PPML Workshop

August 17, 2025

Sam Buxbaum

Lucas M. Tassis

Lucas Boschelli

Giovanni Comarela

Mayank Varia

Mark Crovella

Dino P. Christenson

Traditional Political Polling

- Data collection takes time
- Data collection is human-intensive
- Poor geographic and temporal coverage

West Virginia 2024 Presidential Election Polls

Harris vs. Trump

Source	Date	Sample	Harris	Trump	Other
Research America	8/30/2024	400 LV ±4.9%	34%	61%	5%

Michigan 2024 Presidential Election Polls

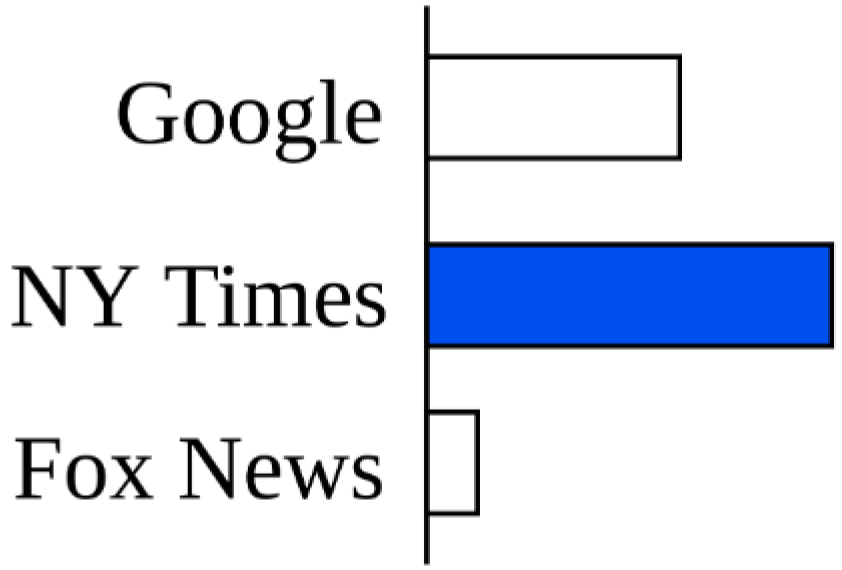
Instantly compare a poll to prior one by same pollster

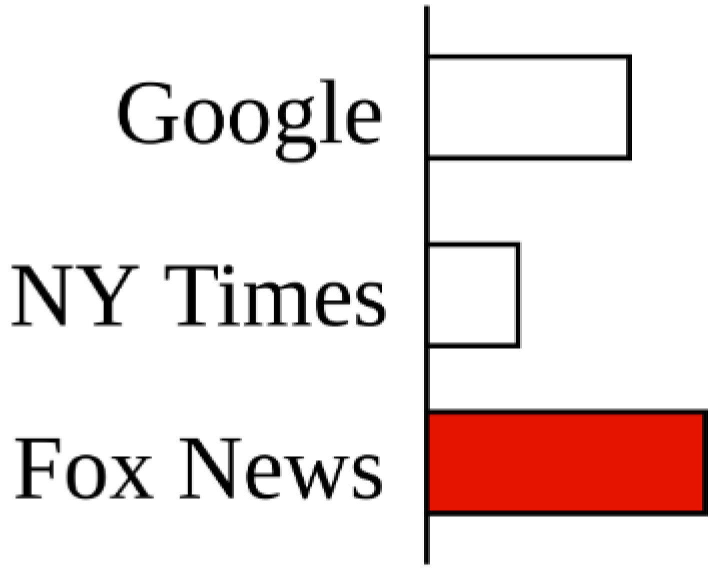
Harris vs. Trump

Source	Date	Sample	Harris	Trump	Other
Average of 23 Polls†			48.6%	46.8%	-
FAU / Mainstreet	11/04/2024	713 LV	49%	47%	4%
Emerson College	11/04/2024	790 LV ±3.4%	50%	48%	2%
Research Co.	11/04/2024	450 LV ±4.6%	49%	47%	4%
InsiderAdvantage	11/03/2024	800 LV ±3.7%	47%	47%	6%
Trafalgar Group	11/03/2024	1,079 LV ±2.9%	47%	48%	5%
MIRS / Mich. News Source	11/03/2024	585 LV ±4%	50%	48%	2%
NY Times / Siena College	11/03/2024	998 LV ±3.7%	47%	47%	6%
Morning Consult	11/03/2024	1,108 LV ±3%	49%	48%	3%
AtlasIntel	11/02/2024	1,198 LV ±3%	48%	50%	2%
Redfield & Wilton	11/01/2024	1,731 LV ±2.2%	47%	47%	6%
The Times (UK) / YouGov	11/01/2024	942 LV ±3.9%	48%	45%	7%
EPIC-MRA	11/01/2024	600 LV ±4%	48%	45%	7%
Marist Poll	11/01/2024	1,214 LV ±3.5%	51%	48%	1%
AtlasIntel	10/31/2024	1,136 LV ±3%	49%	49%	2%
Echelon Insights	10/31/2024	600 LV ±4.4%	48%	48%	4%
MIRS / Mich. News Source	10/31/2024	1,117 LV ±2.5%	47%	49%	4%
UMass Lowell	10/31/2024	600 LV ±4.5%	49%	45%	6%
Washington Post	10/31/2024	1,003 LV ±3.7%	47%	46%	7%
Fox News	10/30/2024	988 LV ±3%	49%	49%	2%
CNN	10/30/2024	726 LV ±4.7%	48%	43%	9%
Suffolk University	10/30/2024	500 LV ±4.4%	47%	47%	6%

Web Browsing for Political Polling

- Can website visits predict political leanings?
- Example - news websites
- More data
- Fully automated





Prior Work

- Web browsing behavior can predict voting results
- Quantifying the 'Comey letter' (Comarela et al.)
- Social media referrals are the best signal

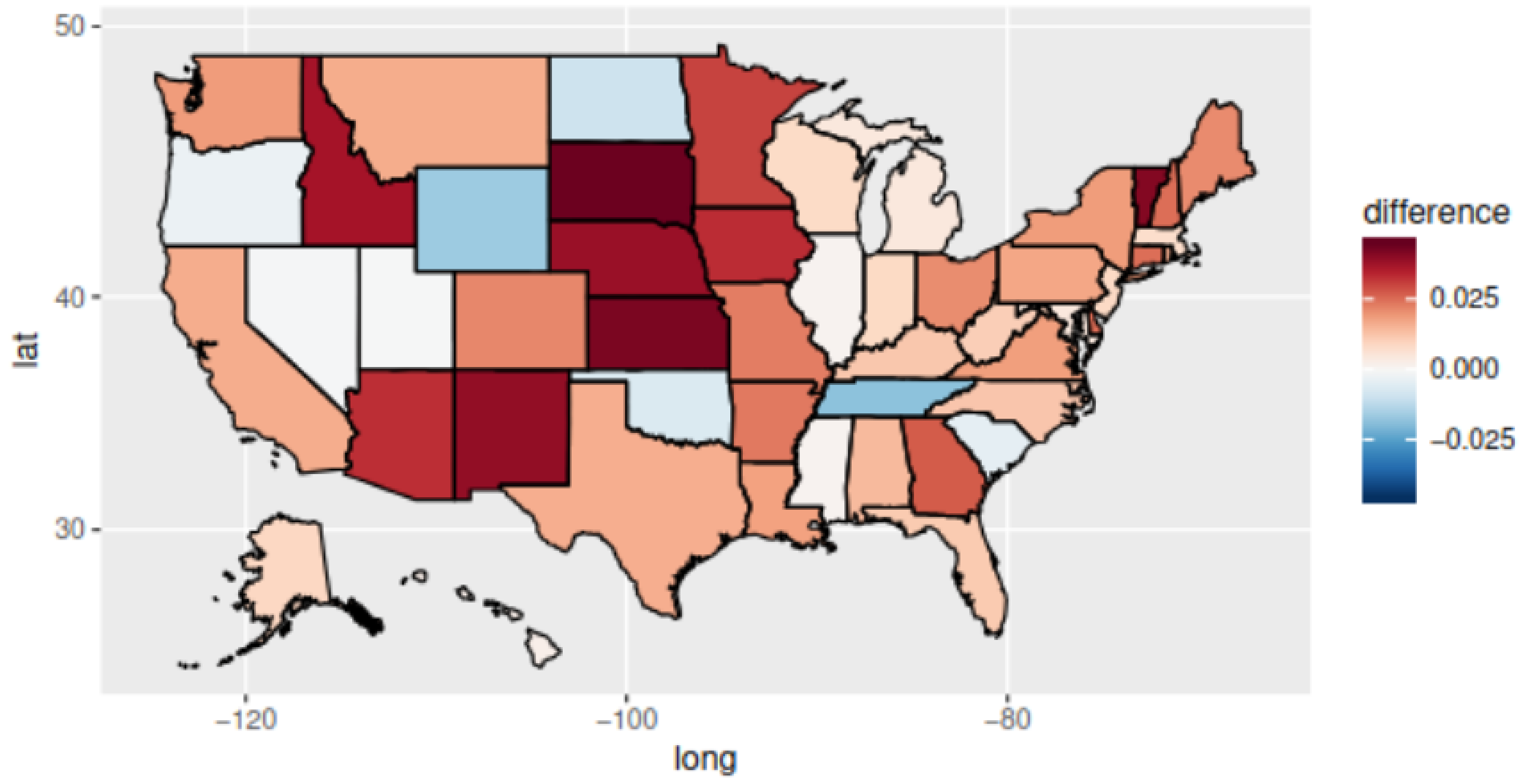


Figure 8: Impact of the 'Comey letter' at the state level.

Two Approaches to Political Polling

Traditional Polling	Web Behavior Analysis
Slow	Immediate
Expensive	Cheap
Coarse-grained insights	Fine-grained insights

VS

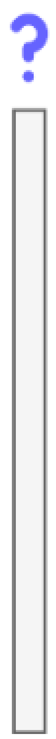
What about privacy?

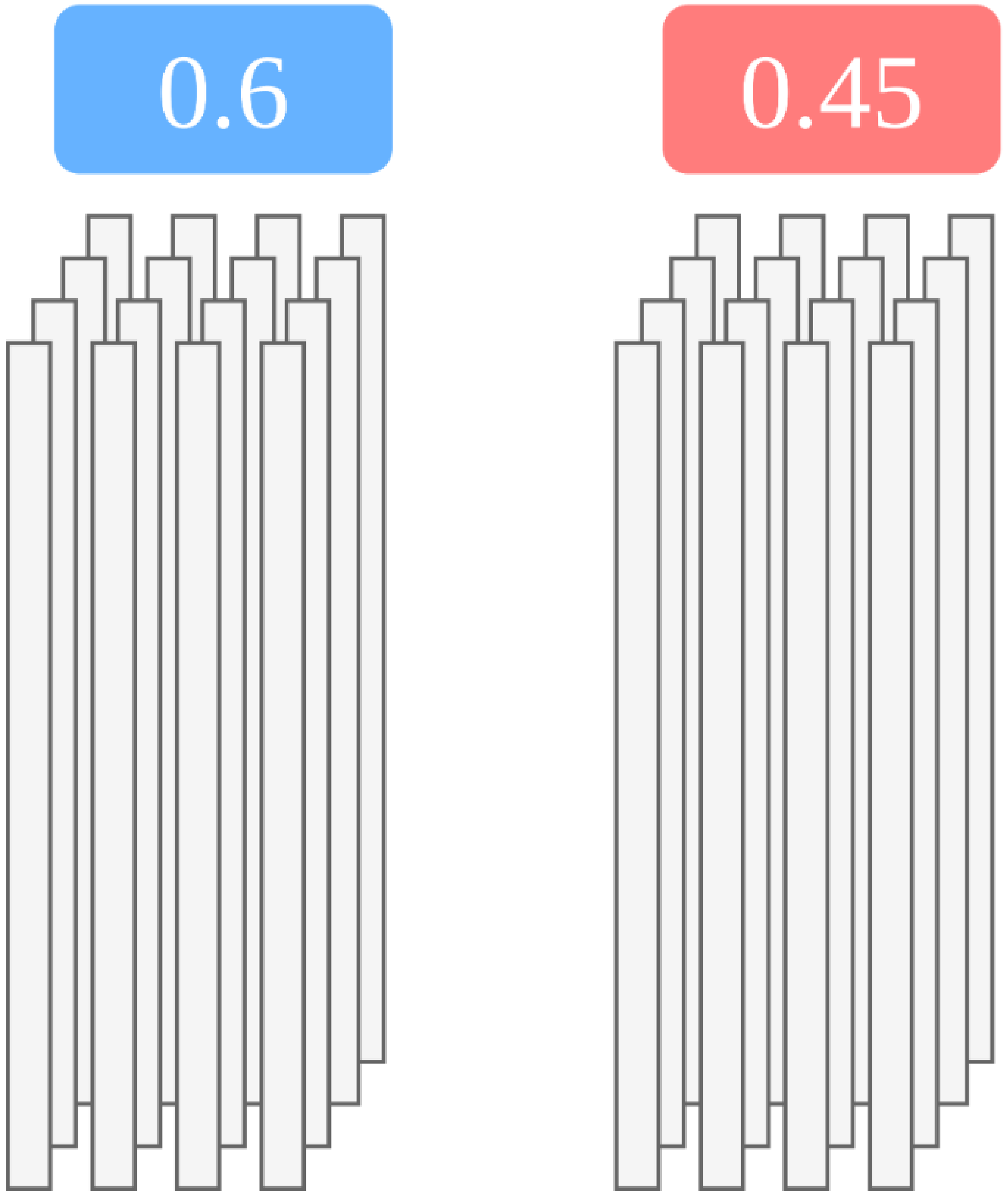
Our Contributions

- We built a system for securely predicting political preferences from web browsing data
- We collected and analyzed data from almost 8000 unique users
- All analysis took place under MPC

Learning from Label Proportions (LLP)

- Each user uploads an *unlabeled* 1,034-element vector every day
  - Number of visits to the top 517 sites
  - Number of times referred to the top 517 sites
- Unlabeled vectors are grouped by state
- Each state has a ground-truth label
- Train on aggregate ground truth
- Predict on an individual level



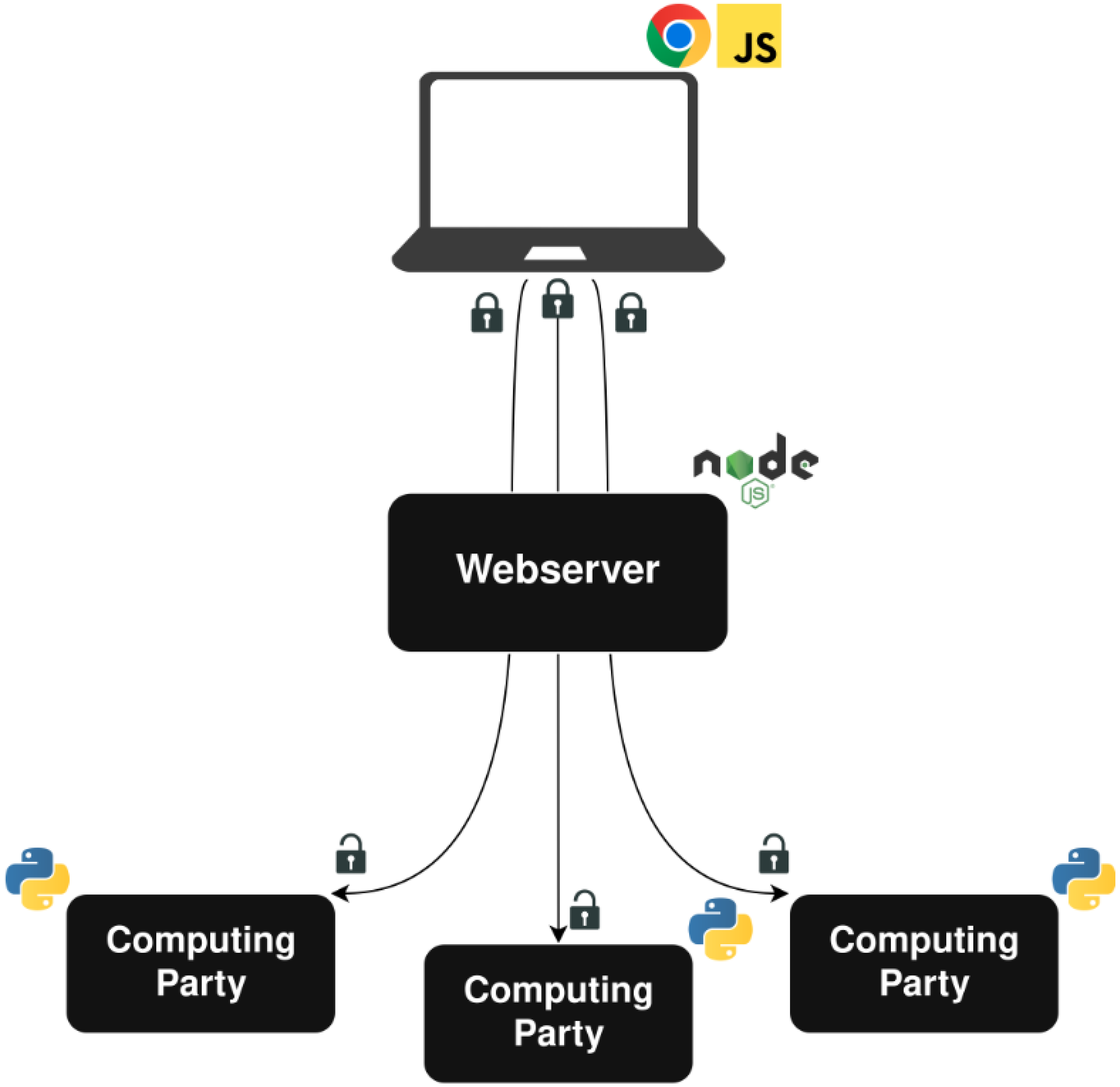


System Design

Client Plugin

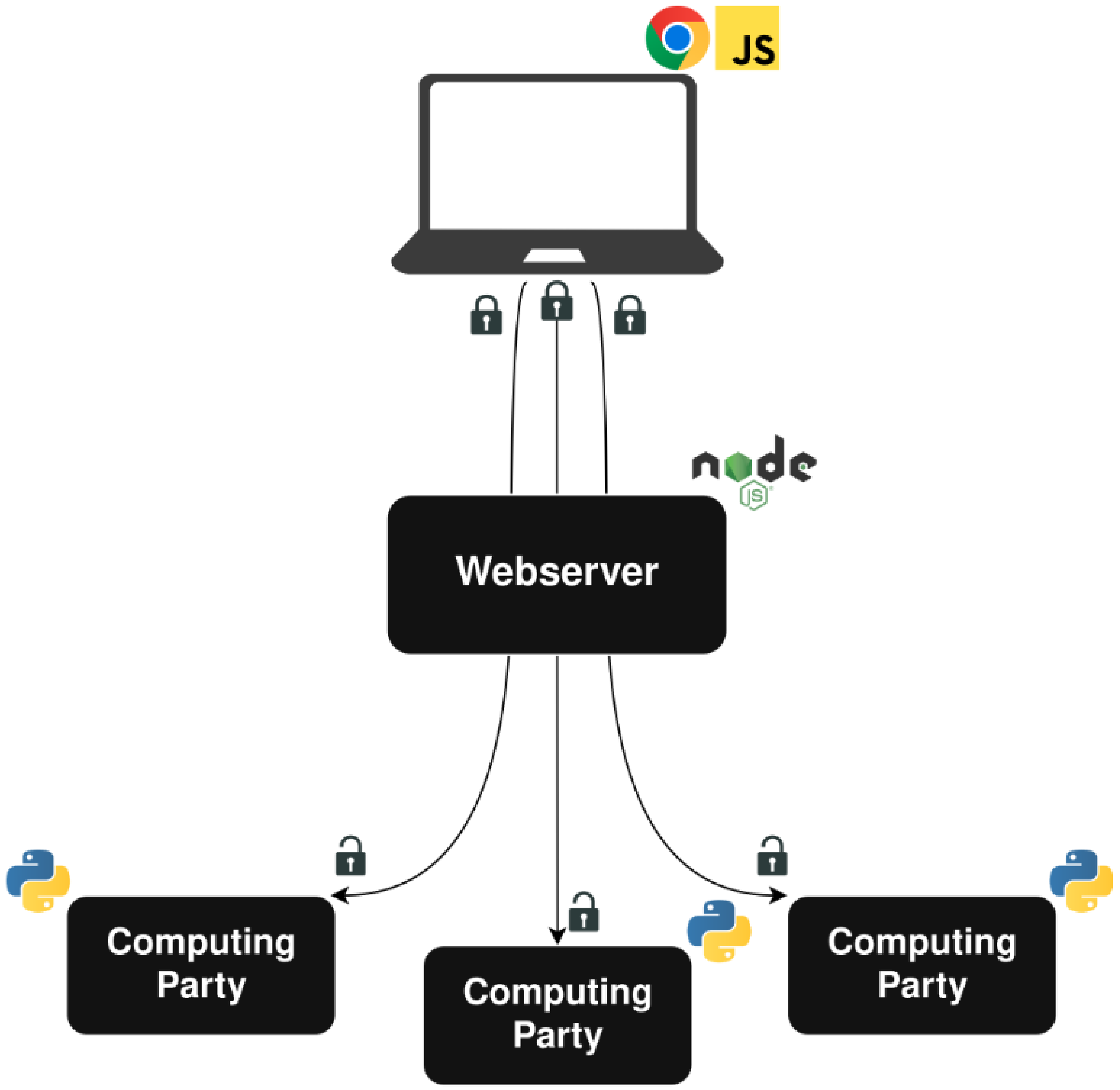
Webserver

MPC Backend



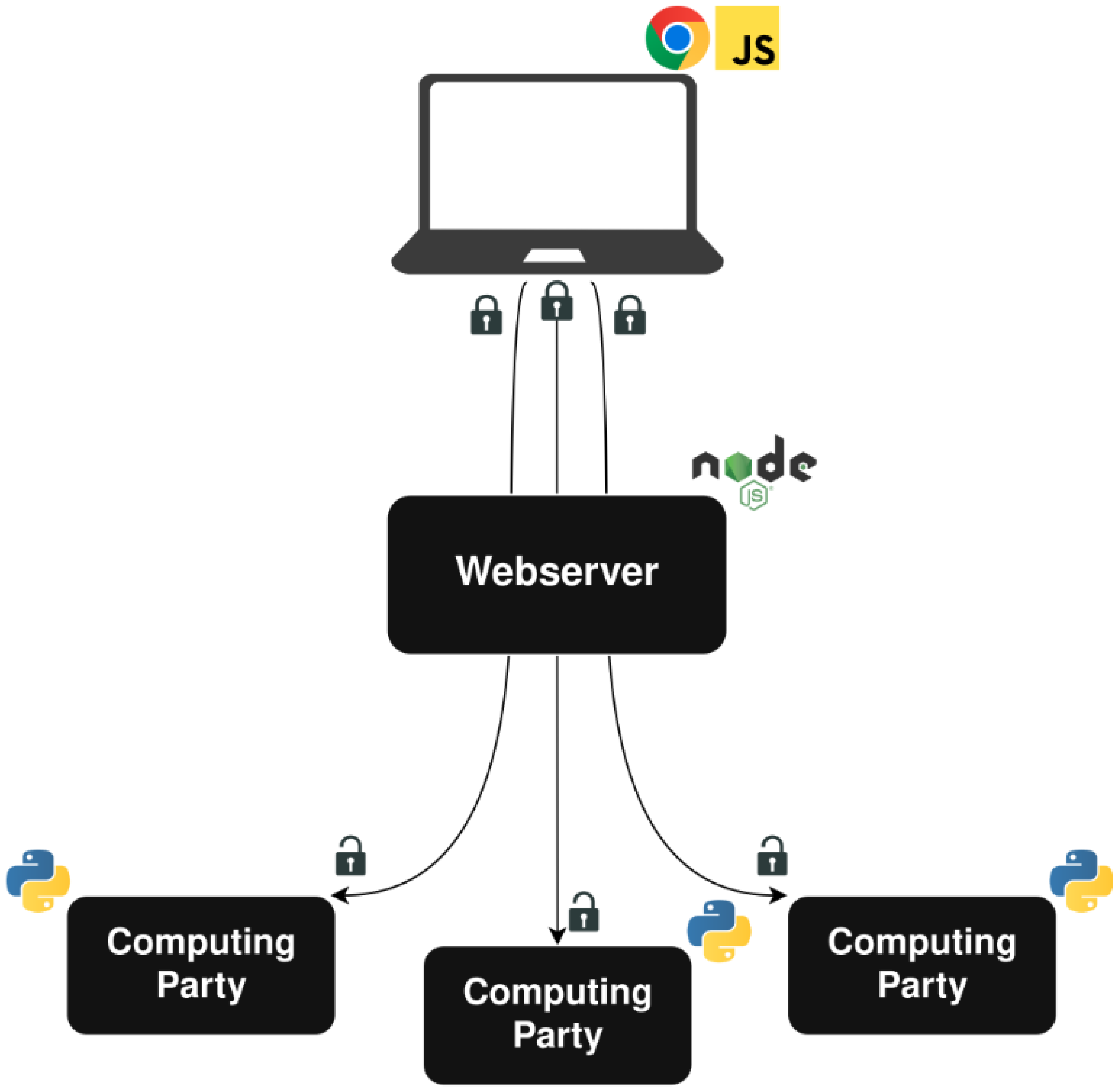
Client Plugin

- Custom-built Chrome plugin to monitor browsing
- Daily data uploads of secret-shared histograms
- Client-side secret sharing and encryption
- Implementation is open source



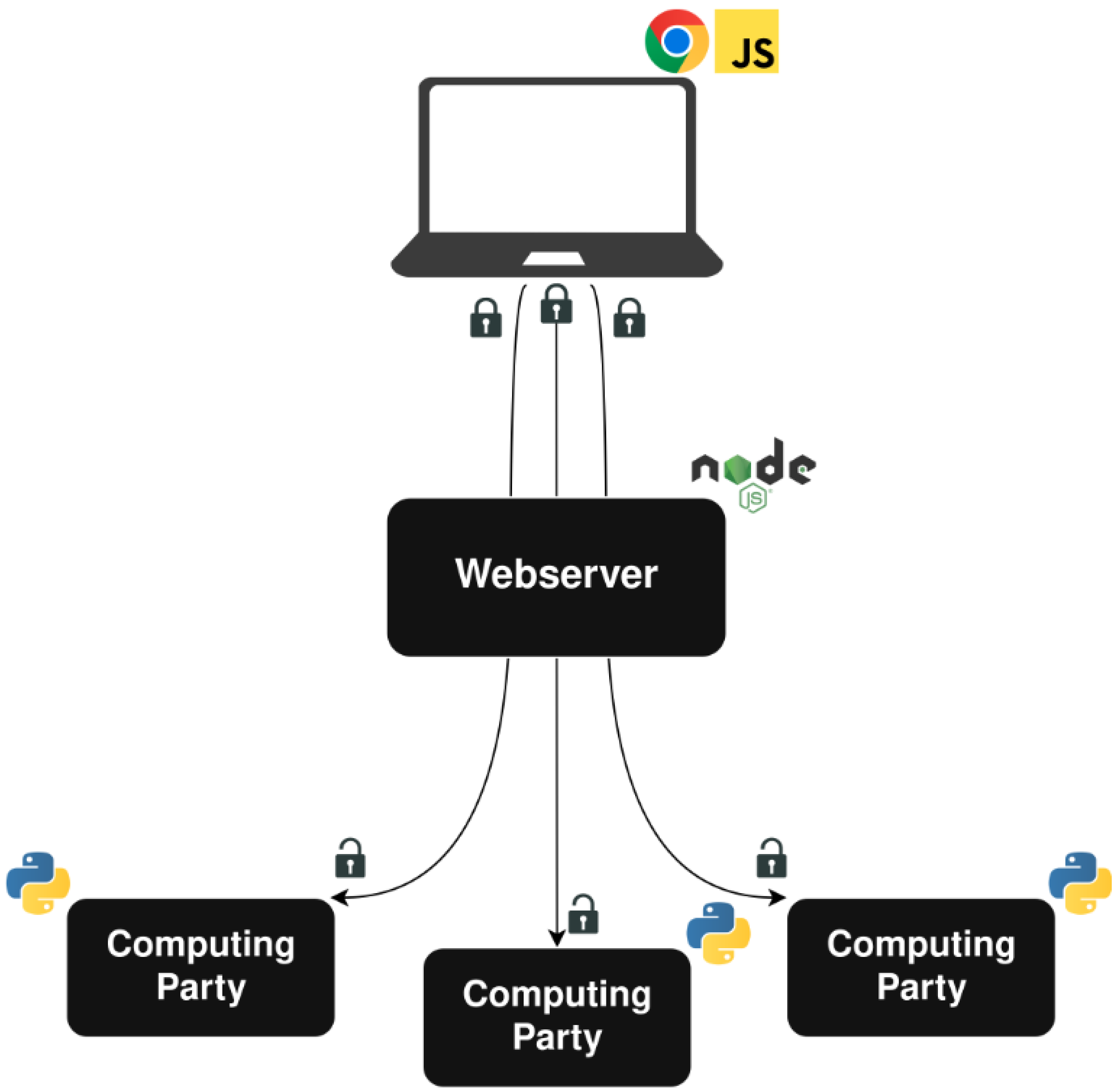
Webserver

- Simplifies interaction with clients
- Collects basic metadata
- Never sees any private data

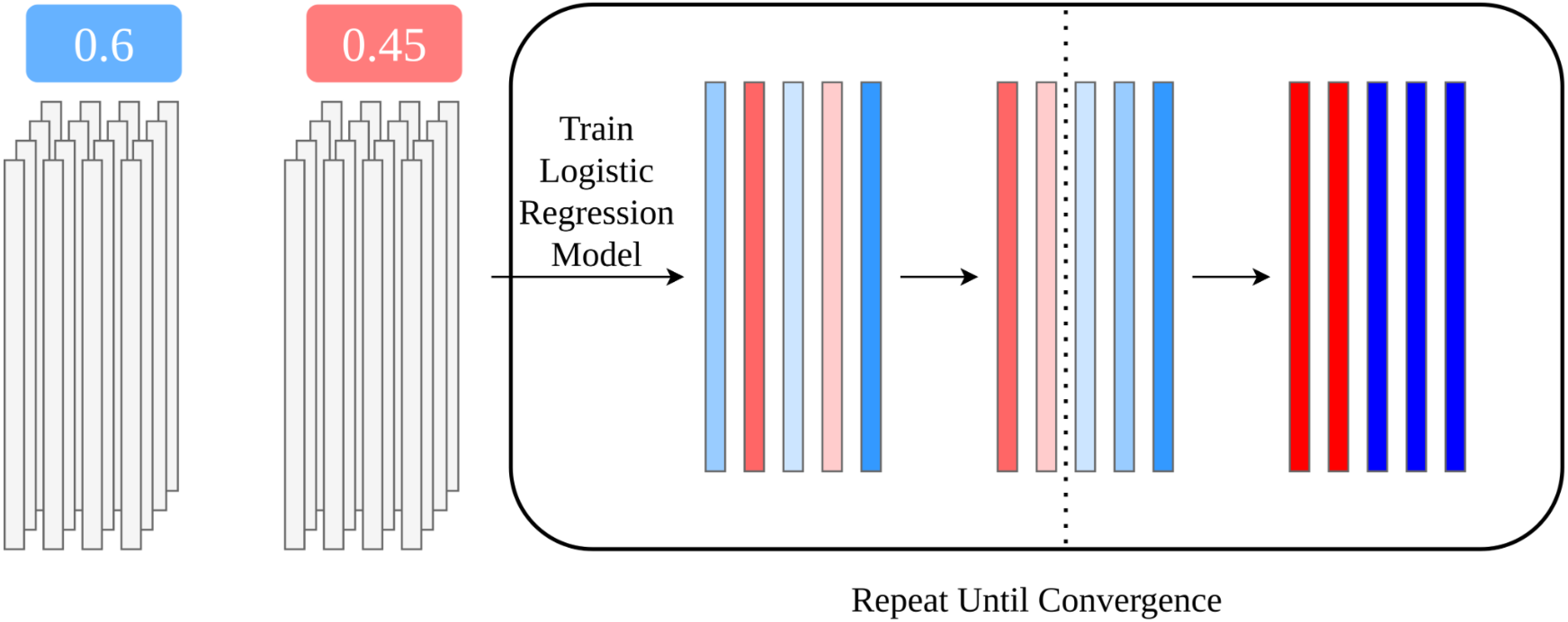


MPC Backend

- We use and augment the CrypTen library
- We implemented an algorithm for LLP under MPC
- Three party computation with an honest majority



The Plaintext Algorithm



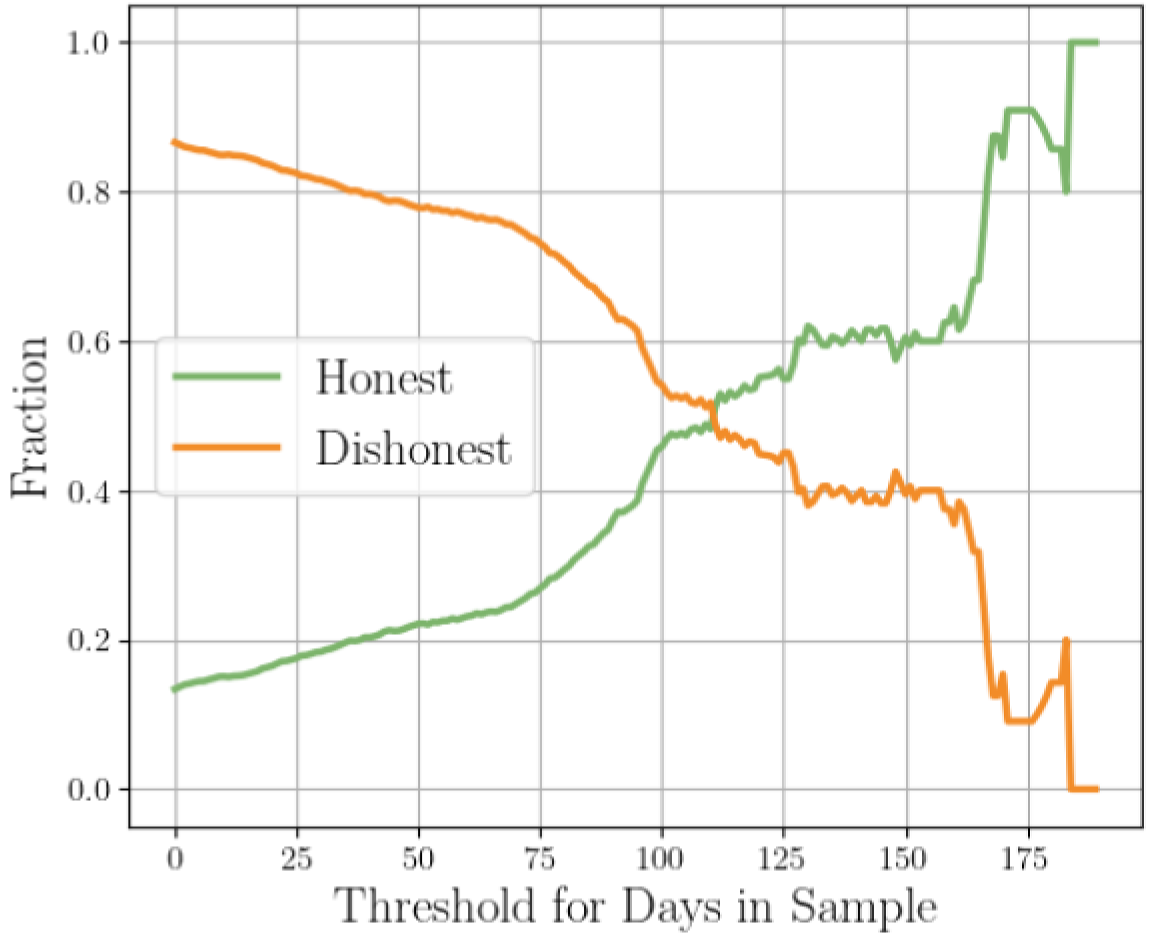
Implementation in MPC

- Initial label assignment can be performed in plaintext
- Training a logistic regression model is supported by CrypTen
- Computing thresholds requires oblivious sorting
- Updated label assignment and convergence checking use secure comparisons
- Practically efficient
- Code will be open source in the future

Lessons Learned and Future Directions

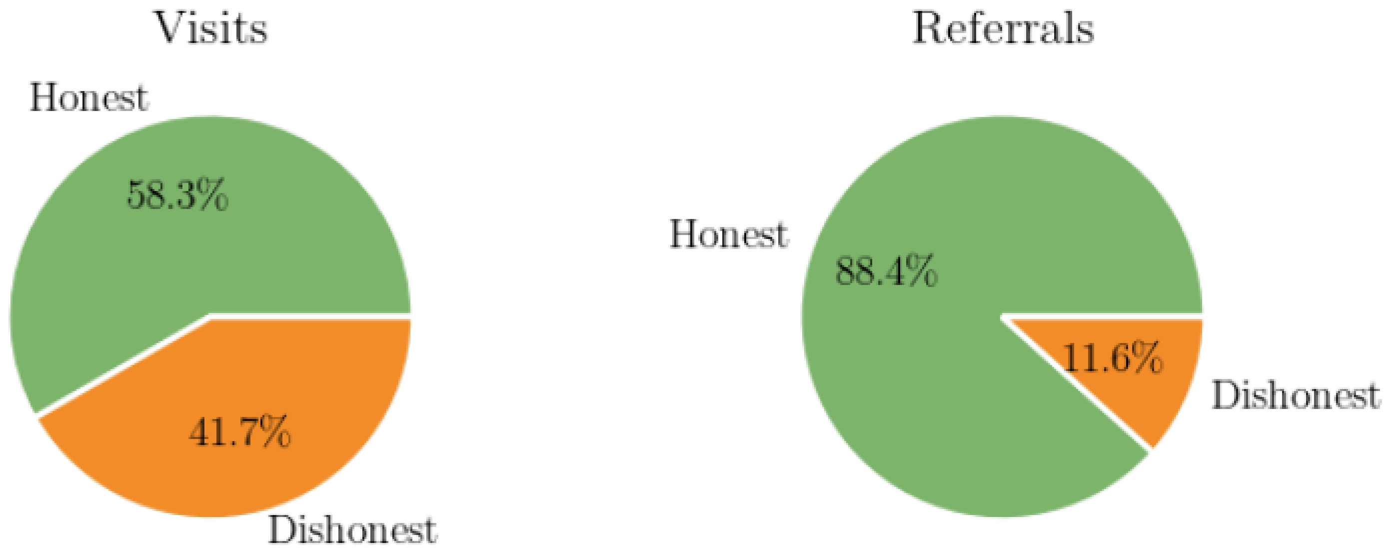
Data Integrity Matters

- Trouble with dishonest reporting of location
- Validation with IP addresses and geolocation
- State-level results are concerning
- Digging deeper on the data
  - Users in the sample for longer are more honest
  - Honest users contribute much richer data



Lesson: Validating and enforcing user honesty should be a priority in future deployments.

Lesson: Our learning process is surprisingly robust to dishonest users.



Strengthening the Threat Model

- AWS as a single point of failure
- Reduce or eliminate trust in the core computation
- Anonymous payments

Thank You!

sambux@bu.edu